基于灰关联分析的 V-MDAV 算法研究 *

张岐山，郑丽君

(福州大学 经济与管理学院，福州 350108)

摘 要：距离度量会影响微聚集算法的聚类效果，为了提高算法的隐私保护能力，采用灰关联分析中的均衡接近度替代 V-MDAV 算法中的欧氏距离度量记录间的距离，提出基于灰关联分析的 V-MDAV 算法，即 V-GRAV(variable-size grey relation to average vector)算法。由于均衡接近度既包含灰关联度对整体接近性的测度，又具有均衡度对序列均衡性测度的特点，克服了欧氏距离受局部奇异值影响较大的问题。因此 V-GRAV 算法在保证信息损失与 V-MDAV 相近的同时，较大程度地降低隐私泄露风险，实验证明算法的有效性。

关键词：隐私保护；V_GRAV 算法；均衡接近度；信息损失；隐私泄露风险

中图分类号：TP301.6 doi: 10.19734/j.issn.1001-3695.2018.06.0464

Research on V-MDAV algorithm based on grey relational analysis

Zhang Qishan, Zheng Lijun

(*School of Economics & Management*, *Fuzhou University*, *Fuzhou* 350108, *China*)

Abstract: Distance measure can affect the clustering effect of microaggregation algorithms, in order to improve the privacy preserving ability of the algorithm, the Euclidean distance in the V-MDAV algorithm are replaced by the balanced adjacent degree in grey relational analysis method to measure the distance between records, and the V-MDAV algorithm based on grey correlation is proposed, called V-GRAV (variable-size grey relation to average vector) algorithm. The balanced adjacent degree includes the characteristic of the measure of grey relational degree to the whole approximation and balanced degree to the sequence balanced degree, which can eliminate the point correlation tendency. It overcomes the problem that the Euclidean distance is greatly influenced by the local singular value. Therefore, the V-GRAV algorithm can reduce the privacy disclosure risk while ensuring that the information loss is similar to V-MDAV algorithm. Our experiments demonstrate that the algorithm is effective.

Key words: Privacy protection; V_GRAV algorithm; the balanced adjacent degree; information loss; privacy disclosure risk

0 引言

随着数据挖掘技术的成熟，利用数据获取有用信息或知识越来越受到学术界及业界的关注。但与此同时，个人的习惯、品味等不愿为人所知的信息都以惊人速度被推断并被消费，因此隐私保护问题日渐突出。

为了解决隐私保护问题，Samarati 等人[1]于 1998 年提出 *k*-匿名技术，该技术要求发布的数据中至少存在 *k* 条不可区分的记录，使攻击者无法通过发布的数据回溯个人，防止隐私信息与个人的匹配。因此，*k*-匿名技术在一定程度上保护了用户的个人隐私。隐私保护在实现过程中，多数采用泛化/隐匿技术[2~5]，而属性值泛化域的确定一直是一个难以解决的问题，而且对数值型数据的泛化易导致语义缺失，造成不必要的信息损失。对此，很多学者将 SDC(statistical disclosure control)技术中的微聚集(Microaggregation)算法用于数据表 *k*-匿名化以解决泛化技术在数值型数据应用上的不足。其中，MDAV 算法是一种实现数值型数据匿名化十分重要且性能较好的的微聚集算法。该算法已被证明在所得到的等价组的同质性方面表现最好。然而 MDAV(maximum distance to average vector)作为固定尺寸的启发式算法，在某些情况下，它会产生远离最优的 k 分组。为此，一种新的启发式多元微聚集方法 V-MDAV(variable-size maximum distance to average vector)[6]应运而生，该方法形成可变尺寸大小的等价组，使分组更自然地适应数据集分布，增加组内同质性。

但是 V-MDAV 算法采用欧氏距离度量记录间的距离，而欧氏距离在度量记录间距离时受奇异值影响较大，因此，V-MDAV 算法在一定程度上会影响隐私保护效果。为此，根据 MDAV 算法基础上引入均衡接近度提出 GRAV 算法[7]的思想，本文提出

收稿日期：2018-06-23；**修回日期：**2018-08-08 **基金项目：**国家自然科学基金青年项目（61300104）；福建省自然科学基金资助项目（2018J01791）

作者简介：张岐山（1962-），男，黑龙江绥化人，教授，博导，博士，主要研究方向为数据挖掘、系统优化与仿真（zhangqs@fzu.edu.cn）；郑丽君（1993-），女，硕士研究生，主要研究方向为隐私保护.

了 V-GRAV 算法，该算法在 V-MDAV 算法基础上，将均衡接近度代替欧氏距离进行 k-聚集。均衡接近度作为灰关联分析的改进方法，其与数理统计中的回归分析等方法不同，它对样本数据量及样本规律性不做限制，而且计算量小，其量化结果与定性分析的结果一致，特别适用于小样本、无明显规律数据的研究[8]。该算法具有灰关联分析方法不受数据分布规律性影响的特点，同时克服了欧氏距离和灰关联分析方法存在的点关联倾向。因此将均衡接近度用于衡量记录间的距离进行 *k*-匿名，能够在保证信息损失量与 V-MDAV 算法相当的同时，较大程度地降低隐私泄露风险。

1 相关工作

1.1 V-MDAV 算法

最优多变量微集聚问题不能在多项式时间内精确求解,因此多变量微聚集是 NP-hard 问题[9]，其唯一可行的多变量微聚集方法是启发式的。MDAV 算法[10,11]是著名的固定尺寸的启发式算法。该算法产生固定基数为 *k* 的等价组，然而，MDAV 作为固定尺寸的启发式算法，在某些情况下，它会产生远离最优的 *k* 分组。

为消除不自然 *k* 分组的影响， Solanas 等人[6]提出 V-MDAV(Variable-size maximum distance to average vector)算法,该算法能适应数据集分布，产生更加自然的分组，增加组内同质性。此外，Solanas 等人提出使用遗传算法[12]微聚集多达 100 条记录的小型多元数据集，提出使用新的 N 元编码来处理微聚集的多变量性质，并进行了一套完整的实验来确定遗传算法主要参数的最佳值，即种群大小，交叉率，变异率等。然而，该方法只适用于小数据集。为了解决该问题，文献[13]将遗传算法与 V-MDAV 算法相结合，利用 V-MDAV 算法对原始大数据集划分成遗传算法可处理的较小子集，而后使用遗传算法来获得微集聚数据集。Huang 等人[14]结合微聚集和曲面细分的优点，提出 Hybrid-VMDAV 算法解决位置隐私问题，同时在 *l*-多样性原则指导下，提出了一种 *l*-多样性的 VMDAV（LD-VMDAV）作为改进，有效地防止时间和空间隐私属性的泄露。

相比 V-MDAV 算法，V-GRAV 算法可以产生更自然的 *k* 分组，两种算法的分组情况见下列例子：

例 1 设 S 为由 9 个具有 2 个属性的记录组成的数据集。

S={(2.4,3),(1.68,4.9),(3.18,5.54),(5.32,3.6),(18.68,11.49),(20.14,9.56),(19.85,10.33),(21.28,10.9),(23,11.5)}

图 1 描绘了当 *k*=3 时，数据集 S 使用 MDAV 算法生成 3 个等价组。其中圆圈表示数据集 S 的记录，三角形表示数据集 S 的整体质心。从图 1 中可以直观看出标有红色的组非常分散，导致整个 3 分组变差。这个例子表明，MDAV 的固定大小的特性可能无法适当地调整生成的 *k* 分组到特定的数据集。图 2 中 V-MDAV 算法根据数据分布情况进行聚类，生成 2 个等价组，很明显，相较图 1 聚类更显合理。由此可得，V-MDAV 算法能够适应微观数据集中记录的自然分布来改进 MDAV 的结果。

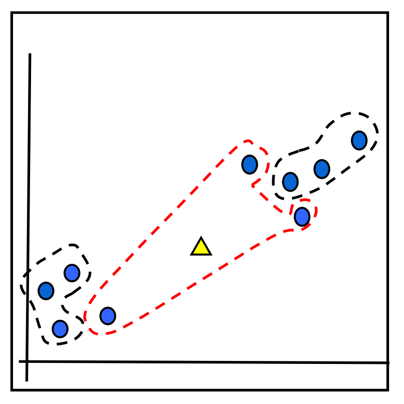


图 1 MDAV 算法运用于数据集 S 时的输出（*k*=3）

Fig.1 output of MDAV algorithm applied to data set S

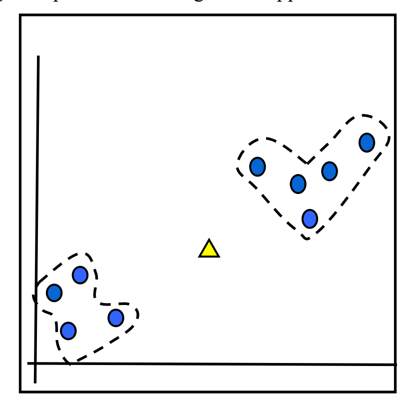


图 2 V-MDAV 算法运用于数据集 S 时的输出(*k*=3)

Fig.2 output of V- MDAV algorithm applied to data set S

1.2 灰关联分析

灰关联分析是邓聚龙教授所提出的灰色系统理论中十分重要的方法，灰关联分析通过序列曲线几何形状的相似程度判断其相关性，曲线越接近则两序列的关联度越大。并且灰关联分析方法对数据量及其分布规律没有要求，其量化结果与定性分析结果也不会出现出入，可用于解决不确定性问题和非线性问题。

传统的灰关联分析法存在着局部点关联倾向，张岐山教授将均衡度引入灰关联度提出均衡接近度[15]的概念，能有效消除点关联倾向，克服传统灰关联分析方法存在的不足。使用均衡接近度作为记录之间的度量方法在聚类算法中已经得到成功应用，文献[16]中利用均衡接近度度量数据间的相似性来克服参数敏感性问题，提高了传统谱聚类算法的性能。李莉琼[17]等人提出一种将均衡接近度结合到 FCM 中的新算法，从全局判断数据间的相似性程度的同时，减弱局部强关联性导致的影响。

定义 1 灰关联度。设 X 为灰关联因子集，$X_0 \in X$ 为参考序列，$X_i \in X$ 为比较序列，$X_0 = \{x_0(k), k \in K\}$，$X_i = \{x_i(k), k \in K\}$，其中 i={1, 2, 3, ...， h}，K={1, 2, 3, ..., n}，

$$r(x_0(k), x_i(k)) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \zeta \max_i \max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \zeta \max_i \max_k |x_0(k) - x_i(k)|} \tag{1}$$

$$r(X_0, X_i) = (1/n)\sum_{k=1}^{n} r(x_0(k), x_i(k)) \tag{2}$$

其中：ξ 为分辨系数，$r(X_0, X_i)$为参考序列 X_0 和比较序列 X_i 的灰关联度。

定义 2 均衡度。给定 R 为灰关联系数序列集，其中，$R_i=\{r(x_0(k),x_i(k)),k\in K\}$，$K=\{1, 2, 3, ..., n\}$，$i=\{1, 2, 3, ..., h\}$，则称

$$p_i(k)=r(x_0(k),x_i(k))/\sum_{k=1}^{n}r(x_0(k),x_i(k)) \tag{3}$$

$$E(X_0,X_i)\overset{\Delta}{=}[-\sum_{k=1}^{n}p_i(k)\cdot\ln p_i(k)]/\ln n \tag{4}$$

$B(R_i)$为参考序列 X_0 和比较序列 X_i 的均衡度，lnn 为灰熵的最大值，同时，由式（3）可知 $p_i(k)$的取值范围为(0, 1)。

定义 3 均衡接近度。设 X 为灰关联因子集，$X_0\in X$ 为参考序列，$X_i\in X$ 为比较序列，$r(X_0, X_i)$为参考序列 X_0 和比较序列 X_i 的灰关联度，$B(R_i)$为参考序列 X_0 和比较序列 X_i 的均衡度，则

$$B(X_0,X_i)=E(X_0,X_i)\times r(X_0,X_i) \tag{5}$$

$B(X_0, X_i)$为参考序列 X_0 和比较序列 X_i 的均衡接近度，均衡接近度越大，则比较序列与参考序列相关性越大。

2　V-GRAV 算法

2.1　算法描述

V-GRAV 算法采用均衡接近度作为记录之间距离的衡量。由于均衡接近度既包含灰关联度对整体接近性的测度，又具有均衡度对序列均衡性测度的特点，因此将均衡接近度引入到微聚集算法中，提出 V-GRAV(variable-size grey relation to average vector)微聚集算法，能为 V-MDAV 微聚集算法的实现方法提供一种新思路。下面详细描述该算法的步骤：

2.1.1 组生成

输入：原始数据集 S，匿名模型尺寸 k，分辨系数 ξ，增益因子 γ。

输出：经过处理后的匿名数据表 S’。

1、计算数据集 S 中所有记录间的均衡接近度 D；

2、计算数据集 S 的质心 C；

3、在未分配记录中找出距离质心 C 均衡接近度最小的记录 r；

4、以 r 为中心，找出与 r 均衡接近度最大的 k-1 个记录，将这些记录组成一个等价组；

5、扩展组（见 2.1.2）；

6、继续步骤 3，直到未分配记录小于 2k；

7、如果剩余记录数小于 k，则将剩余记录分配到其最近的子集；否则，剩余记录形成最后一个子集。

2.1.2 扩展组

扩展步骤允许 V-GRAV 适应记录的自然分布。在生成 k 条记录的子集之后，扩展步骤找到可能加入子集的候选记录，并且如果这些候选记录中的任何一个比其他未分配的记录更接近子集，则将它们添加到子集中。扩展步骤的工作原理如下：

给定具有 m 条记录的组 g，最接近组 g 的未分配记录 e_{max}，e_{max} 与组 g 的最大均衡接近度 b_{in} 以及 e_{max} 与未分配记录的最大均衡接近度 b_{out}，公式如下：

$$b_{in}=\max_{j\in[1,N_{un}]}\max_{i\in[1,m]}b(e_i^g,e_j) \tag{6}$$

$$e_{\max}=\arg\max_{j\in[1,N_{un}]}\max_{i\in[1,m]}b(e_i^g,e_j) \tag{7}$$

$$b_{out}=\max_{j\in[1,N_{un}],e_{\max}\neq e_j}b(e_{\max},e_j) \tag{8}$$

其中：$eg\ i$ 表示组 g 中的第 i 条记录，e_j 表示未分配数据集的第 j 条记录，N_{un} 表示未分配数据集的数量。如果满足等式（7）的记录数不止一条，则随机选取其中一条记录作为 e_{max}。

最后，判断是否要将 e_{max} 加入组 g 中，需进行 b_{in} 与 b_{out} 的比较。式（9）给出了判定标准：

$$Add_Record=\begin{cases}YES\ \ if\ \ \gamma\cdot b_{in}>b_{out}\\ NO\ \ otherwise\end{cases} \tag{9}$$

为了提高 V-GRAV 的适应性，必须调整增益因子 γ。然而增益因子最佳值的确定并不简单，而且由于篇幅限制，本文不加以讨论。重复扩展过程直到组大小等于 2k - 1 或表达式（9）中的条件不被满足，因为它在文献[11]中表明，最佳的 k 分组为每个组包含 k 和 2k-1 个记录。

通过对比发现 V-GRAV 算法与 V-MDAV 算法主要区别于以下几个方面：

a）V-GRAV 算法通过均衡接近度来度量记录间的相似性，而 V-MDAV 算法采用欧氏距离来测度记录间的距离。均衡接近度综合考虑了熵关联度和点关联度，因此既包含了对序列间点的距离接近性的测度，又包含了对整体的无差异性接近的测试。

b）V-GRAV 算法中的均衡接近度度量的是记录间的相似性，均衡接近度越大，两条记录越相似，距离越接近，反之则越远。而 V-MDAV 算法中的欧氏距离测度的距离越大，相距越远，与均衡接近度刚好相反。

2.2　算法评估

本文提出 V-GRAV 算法并将其应用于隐私保护中，数据发布隐私保护要求算法在达到隐私保护目的的同时，要保证数据的可用性。V-GRAV 算法实现了匿名化技术的同时，由于通过等价组质心替代组内记录，会产生数据失真，降低了数据的可用性。因此，衡量本文提出的 V-GRAV 算法的性能表现，需要从匿名表的信息损失度和隐私泄露风险两个角度出发来评价微聚集算法的有效性。

2.2.1 匿名表信息损失度

计算信息损失的方法有很多，文献**错误!未找到引用源。**中的度量方法就是一种度量连续型数据的常用方法。本文引用该文献中的 IL（information loss）来表示信息损失，计算一个匿名数据表的信息损失度的过程如下：

1）计算原始数据表的总体同质性测度和 SST

假设原始数据表有 n 条记录数，数据表经过匿名划分为 h

个等价组，每个等价组的记录数为 m_i，则 SST 计算过程如下：

$$\overline{W}=\frac{1}{n}\sum_{l=1}^{n}W_l \tag{10}$$

$$SST=\sum_{i=1}^{h}\sum_{j=1}^{m_i}d(W_{ij}-\overline{W}) \tag{11}$$

其中，W_l 为原始数据表中的第 l 条记录，式（10）则计算整个原始数据表的均值。式（11）中的 W_{ij} 表示匿名数据表中第 i 个等价组里的第 j 条记录。SST 将匿名表中的所有记录与原始数据表均值的差值求和，得出原始数据表的总体同质性测度和。

2）计算匿名化后等价类的同质性测度和 SSE

$$\overline{W_i}=\frac{1}{m_i}\sum_{j=1}^{m_i}W_{ij} \tag{12}$$

$$SSE=\sum_{i=1}^{h}\sum_{j=1}^{m_i}d(W_{ij}-\overline{W_i}) \tag{13}$$

其中，式（12）计算每个等价组的均值，并将每个等价组的记录与该等价组的均值进行差值计算求和得出匿名化后等价类的同质性测度和 SSE。

3）计算匿名化后的总信息损失度 IL

$$IL=\frac{SSE}{SST} \tag{14}$$

由上述可知，对于一个给定的静态数据集，在 IL 评价模型下，其总体同质性 SST 是固定值，而 SSE 值则因 k-划分情况不同而发生变化。可见，SSE 在 IL 评判算法的优劣中占主导地位，若等价组的组内同质性越大，其 SSE 值就越小，匿名后信息损失度就越小。

2.2.2 匿名表隐私泄露风险

概率记录链接方法（probabilistic record linkage）和基于距离的记录链接隐私泄露风险评价模型 DLD（distance linked disclosure risk）是评价隐私泄露风险最广泛使用的方法[19]，由于基于距离的记录链接方法更易于实施和操作，因此本文采用 DLD 模型进行 V-GRAV 算法隐私泄露风险的评价。

DLD 模型计算原始记录和受保护记录之间的距离，其使用记录链接来反映匿名记录在多大程度上可以被重新识别。假设给定含有 n 个记录的原始数据集 $S(s_1, s_2, \ldots, s_n)$，其匿名后数据集为 S'（$s_1', s_2', \ldots, s_n'$），对于匿名数据集 S'中的记录 s_i'，计算出原始表 S 中距离该记录最近两个记录，如果这两个记录中含有 s_i'匿名前的真实记录 s_i，则称元组 s_i' 链接成功。

假设匿名表中能够链接成功的记录数为 linked_records，而匿名表中记录总数为 total_records，那么隐私泄露的度量为：

$$DLD=linked_records/total_records \tag{15}$$

3　实验

3.1　实验环境

3.1.1 数据集

本文使用 Tarragona、Census 和 EIA 三组经典的数据集作为实验的数据集。其中，Tarragona 数据集为 1995 年塔拉戈纳地区 834 家企业的信息；Census 数据集为 2000 年美国统计局提供的人口普查信息；EIA 数据集为 1996 年美国能源信息管理局提供的美国能源信息。三个数据集分别包含 12、12、9 个数值型属性和一个敏感属性以及 834、1080、4092 条记录数。

3.1.2 实验软/硬件环境

硬件环境：Inte Core 3.30 GHz CPU，4096 MB RAM，Windows 7 32 位操作系统。

编程环境：Mathworks Matlab R2014a。

3.2　信息损失度分析

采用式（14）计算出 V-MDAV 算法与 V-GRAV 算法在不同 k 值下的信息损失度，实验结果如图 3 所示。其中，V-MDAV 算法的实验结果用实线表示，本文提出的算法 V-GRAV 用虚线表示，本文增益因子取 0.2，分辨系数取 1.8，增益因子和分辨系数的确定较为复杂，由于篇幅限制，本文不作讨论。

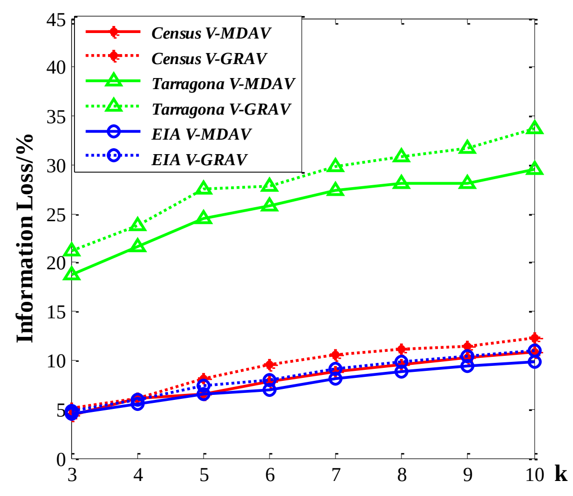


图 3　V-MDAV 与 V-GRAV 算法 k-匿名化信息损失比较

Fig.3　Loss comparison of k-anonymity information between V-MDAV and V-GRAV algorithm

图 3 中，k 表示等价组的最小尺寸，随着 k 的增大，信息损失度呈上升趋势，这是由于等价组扩张，组内的记录数增加使得组内同质性减小，致使信息损失不断上升。

对于不同的数据集，由图中可得三个数据集中 Tarragona 的信息损失度最大，Census 其次，EIA 最小。其中三个数据集的记录数关系为：Tarragona < Census < EIA，在 k 值相同的情况下，数据集越大，在划分等价组时候可选的记录数越多，更易于选取同质性更高的记录，使得组内同质性更高，因此 EIA 数据集匿名后的信息损失度更小。

同时，对比 V-MDAV 和 V-GRAV 两种算法发现：V-GRAV 算法的信息损失度高于 V-MDAV 算法。由于在信息损失度计算过程中，采用的是欧氏距离进行度量，更符合 V-MDAV 算法等价组划分过程，由此 V-MDAV 算法的信息损失度更低，但两种算法间的信息损失度差距幅度不超过 5%，因此 V-GRAV 的数据可用性得以保证。

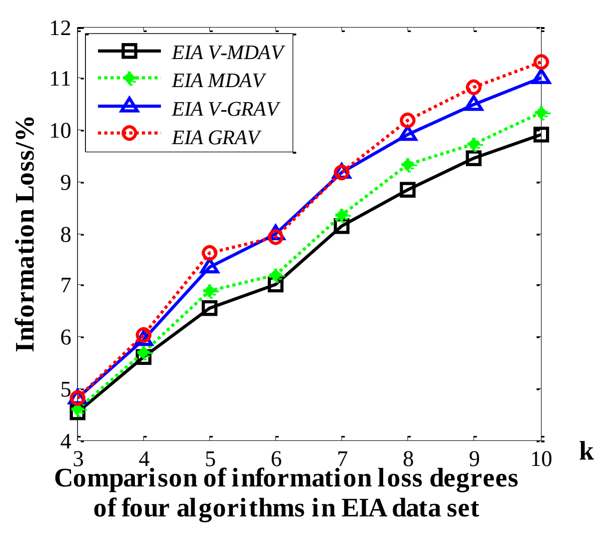


图 4　EIA 数据集下 4 种算法 k-匿名化信息损失比较

Fig.3　Loss comparison of *k*-anonymity information of four algorithm in ELA data set

由于三种数据集信息损失度趋势相同，故图 4 选择其中一种数据集（EIA）用于分析。由图 4 可知,可变尺寸的微聚集算法信息损失度小于固定尺寸的微聚集算法的信息损失度。这是由于可变尺寸的微聚集算法能够适应微观数据集中记录的自然分布情况进行聚类，产生更合理的聚类结果。

并且两种 GRAV 算法的信息损失度高于两种 MDAV 算法，同样是由于在信息损失度计算过程中，采用的是欧氏距离进行度量，更符合 MDAV 算法等价组划分过程，由此两种 MDAV 算法的信息损失度更低。

3.3　隐私泄露风险分析

采用式（15）计算出 V-MDAV 算法和 V-GRAV 算法的隐私泄露风险，见图 5。图 5 的(a)~(c)中分别展示了在 Census、Tarragona、EIA 3 个数据集下 2 种算法的隐私泄露风险随 *k* 值变化的情况。实线为 V-MDAV 算法，虚线为 V-GRAV 算法。

图 5(a)~(c)显示了 3 种数据集下隐私泄露风险趋势相同，并且随着 k 的增加，隐私泄露风险呈下降趋势，由于 k 增加，对于给定的静态数据集，随着等价类内元组的增加，类内同质性减小，信息失真度大，攻击者回溯用户身份的可能性降低，使得隐私泄露风险下降。同时可以看到整体的隐私泄露风险：Tarragona < Census < EIA。因为数据可用性和隐私泄露风险是对立的概念，数据可用性越高，隐私泄露风险越大。

最后，对比 V-MDAV 算法，V-GRAV 算法实现更低的隐私泄露风险度。这是由于 DLD 评价模型在计算链接距离时直接采用欧氏距离公式，与 V-GRAV 算法划分等价类的标准不一致。综合以上原因，用 V-GRAV 算法匿名后的数据集通过 DLD 模型往往链接不到的真实记录，隐私泄露风险较低。

由于三种数据集隐私泄露风险趋势相同，故图 6 选择其中一种数据集（EIA）用于分析。

对比图 6 中 4 种算法，两种 GRAV 算法的隐私泄露风险低于两种 MDAV 算法，同样是由于 DLD 评价模型与两种 GRAV 算法划分等价类的标准不一致。

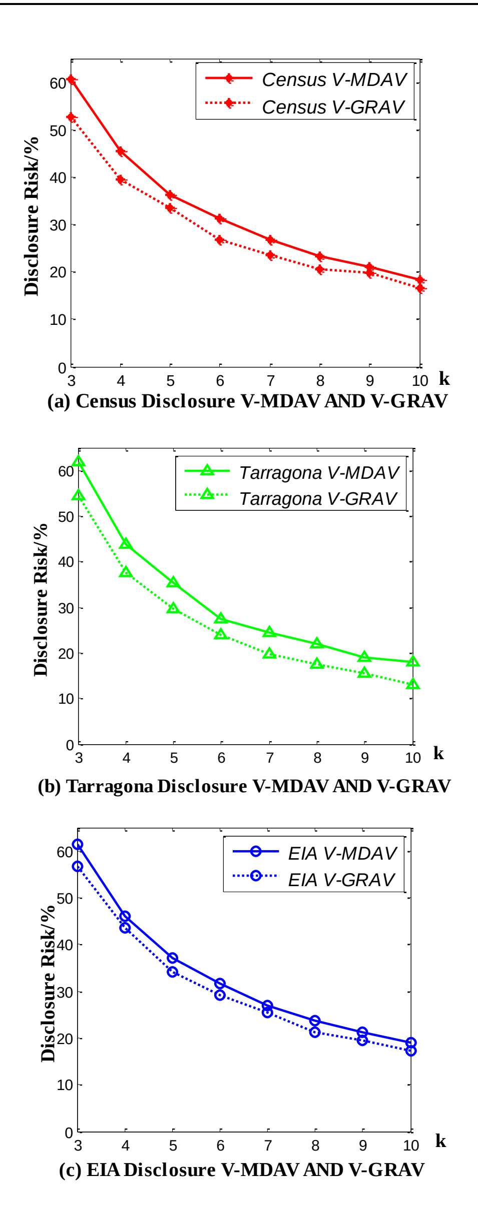


图 5　V-MDAV 与 V-GRAV 算法 *k*-匿名化隐私泄露风险比较

Fig.5　Privacy leakage risk comparison of k-anonymity information between V-MDAV and V-GRAV algorithm

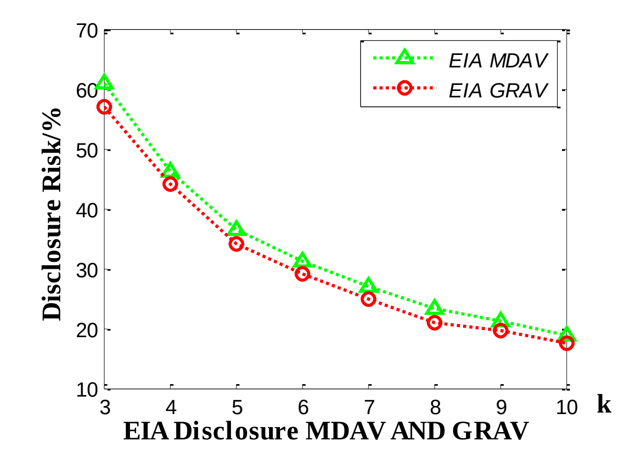


图 6　EIA 数据集下 4 种算法 *k*-匿名化隐私泄露风险比较

Fig.6　Privacy leakage risk comparison of k-anonymity information of four algorithm in EIA data set

同时，由图 4 和 6 可知，可变尺寸的微聚集算法隐私泄露风险与固定尺寸的微聚集算法相差无几情况下，可变尺寸的微

聚集算法信息损失度小于固定尺寸的微聚集算法的信息损失度。可证明可变尺寸的微聚集算法的有效性。

3.4 V-GRAV 算法综合评价

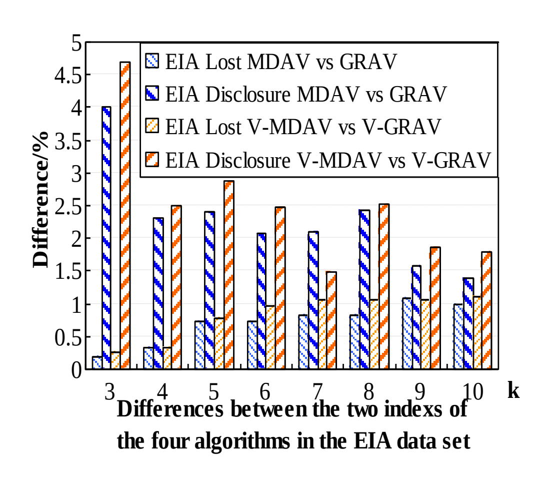


图 7 EIA 数据集下 4 种算法在 2 个指标差值比较

Fig.7 Differences between two indexs of four algorithms in ELA data set

图 7 中，4 种算法两两计算信息损失度差值和隐私泄露风险差值（取正值）。其中，蓝色模块表示固定尺寸微聚集算法 2 个指标的差值，橙色模块表示可变尺寸微聚集算法 2 个指标的差值。

对比 MDAV 和 GRAV 两种算法发现：MDAV 和 GRAV 算法的信息损失度差值不超过 1.5%，而隐私泄露风险的差值基本超过 1.5%，信息损失度差值小于隐私泄露风险差值，说明，均衡接近度引入 MDAV 算法所提出的 GRAV 算法相比 MDAV 算法，在损失较少的信息可用性情况下，较大程度得提升了隐私保护能力，较好地保护了用户的隐私，论证了将均衡接近度引入微聚集算法的有效性。

同样，对比 V-MDAV 和 V-GRAV 两种算法发现：V-GRAV 算法的隐私泄露风险明显低于 V-MDAV 算法。且 V-MDAV 和 V-GRAV 算法的信息损失度差值不超过 1.5%，而隐私泄露风险的差值基本超过 1.5%，信息损失度差值小于隐私泄露风险差值，由此可知，本文提出的 V-GRAV 算法能较大程度降低隐私泄露风险。

综上所述：本文将均衡接近度作为距离度量提出的 V-GRAV 算法是行之有效的。

3.5 实验小结

根据实验结果可以得出如下结论：

a)由 MDAV 算法和 GRAV 算法可证均衡接近度引入微聚集算法的有效性。

b)V-GRAV 算法能够在保证信息损失度情况下降低隐私泄露风险，说明采用均衡接近度作为距离度量方式的微聚集算法是行之有效的。

c)随着 k 值增大，V-GRAV 算法产生的匿名表的信息损失度增大，隐私泄露风险减小。

d)与 V-MDAV 算法一致，随着数据集中记录数量的增大，V-GRAV 算法产生的匿名表信息损失度有下降的趋势。

e)信息损失度与隐私泄露风险呈对立关系，信息损失度越大，隐私泄露风险越小。

总之，由实验结果可知：V-GRAV 算法可用于实现 k-匿名模型并能得到具有较强隐私保护能力的匿名表。

4 结束语

距离度量是微聚集算法在进行 k-聚类的关键问题，针对欧氏距离在距离度量过程中易受奇异值影响，存在点关联倾向的问题，本文根据将均衡接近度引入 MDAV 算法能提高算法隐私保护能力的思想，提出一种新的多元微聚集方法 V-GRAV 算法。V-GRAV 算法采用均衡接近度取代 V-MDAV 算法中的欧氏距离进行记录间距离的度量，均衡接近度包含整体接近性的测度和序列均衡性测度，能消除点关联倾向，使得 V-GRAV 算法能够在保证信息损失度与 V-MDAV 算法接近的情况下降低隐私泄露风险。然而本文的研究尚存在不足之处，计划在未来研究并改进以下几点：a)详细分析增益因子确定给定数据集的最优值;b)本文提出的算法以均衡接近度为度量方式，而均衡接近度适用于数值型数据，因此 V-GRAV 算法不适用于分类型属性，在今后的研究中，需要考虑如何使用灰关联方法实现混合型数据的匿名化。

参考文献：

[1] Samarati P, Sweeney L. Generalizing data to provide anonymity whe-n disclosing information (abstract) [C]// Proc of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems. New York: ACM Press, 1998: 188.

[2] Valls A. Semantic adaptive microaggregation of categorical microda-ta [J]. Computers & Security, 2012, 31 (5): 653-672.

[3] Li Jiuyong, Baig M M, Sattar S A H M, *et al.* A hybrid approach to prevent composition attacks for independent data releases [J]. Information Sciences, 2016, 367-368: 324-336.

[4] Amiri F, Yazdani N, Shakery A. Bottom-up sequential anonymization in the presence of adversary knowledge [J]. Information Sciences, 2018, 450: 316-335.

[5] Li Boyu, Liu Yanheng, Han Xu, *et al.* Cross-Bucket generalization f or information and privacy preservation [J]. IEEE Trans on Knowledge & Data Engineering, 2018, 30 (3): 449-459.

[6] Solanas A, Martínez-Ballesté A. V-MDAV: a multivariate microaggr-egation with variable group size [J]. Seventh Compstat Symposium of the Iasc, 2006.

[7] 郑群华. 基于灰关联分析的隐私保护 k-匿名模型及算法研究 [D]. 福州: 福州大学, 2012. (Zheng Qunhua. Research on k-anonymity Model and Algorithm for Privacy Preserving based on Grey Relational Analysis [D]. Fuzhou: Fuzhou University, 2012.)

[8] 刘思峰, 杨英杰, 吴利丰. 灰色系统理论及其应用 [M]. 7 版. 北京: 科学出版社, 2014: 63-108. (Liu Sifeng, Yang Yingjie, Wu Lifeng. Grey system theory and its application [M]. 7th ed. Beijing: Science Press, 2014:

63-108.)

[9] Laszlo M, Mukherjee S. Iterated local search for microaggregation [J]. Journal of Systems & Software, 2015, 100: 15-26.

[10] Rebollo-Monedero D, Forné J, Pallarès E, *et al.* A modification of the lloyd algorithm for k-anonymous quantization [J]. Information Sciences, 2013, 222 (2): 185-202.

[11] Soria-Comas J, Domingo-Ferrer J. Differentially private data publis-hing via optimal univariate microaggregation and record perturbation [J]. Knowledge-Based Systems, 2018, 153: 78-90.

[12] Solanas A, GonzáLez-NicoláS U, Martinez-Balleste A. A variable-MDAV-based partitioning strategy to continuous multivariate micr-oaggregation with genetic algorithms [C]// Proc of International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 2010: 1-7.

[13] Solanas A, González-Nicolás Ú, Martínez-Ballesté A. Mixing g-enetic algorithms and V-MDAV to protect microdata [J]. Computational Intelligence for Privacy and Security, 2012, 394: 115-133.

[14] Huang Kuanlun, Kanhere Salil S, Hu Wen. Preserving privacy in par-ticipatory sensing systems [J]. Computer Communications, 2010, 33 (11): 1266-1280.

[15] 张岐山，邓聚龙，邵勇．均衡接近度灰关联分析方法 [J]. 华中理工大学学报, 1995, 23 (11): 94-98. (Zhang Qishan, Deng Julong, Shao Yong. A grey correlational analysis by the method of balance and approach [J]. Journal of Huazhong University of Science and Technology, 1995, 23 (11): 94-98.)

[16] 郭昆，张岐山．基于灰关联分析的谱聚类 [J]. 系统工程理论与实践, 2010, 30 (7): 1260-1265. (Guo Kun, Zhang Qishan. Spectral clustering based on grey relational analysis [J]. System Engineering Theory and Practice, 2010, 30 (7): 1260-1265.)

[17] 李莉琼，刘漳辉，郭昆．基于灰关联分析的模糊 C 均值算法 [J]. 福州大学学报：自然科学版, 2016, 44 (2): 170-175. (Li Liqiong, Liu Zhanghui, Guo Kun. Fuzzy C-means based on grey relational analysis [J]. Journal of Fuzhou University: Natural Science Edition, 2016, 44 (2): 170-175.)

[18] Reza Mortazavi, Saeed Jalili. Preference-based anonymization of nu-merical datasets by multi-objective microaggregation [J]. Information Fusion, 2015, 25 (C): 85-104.

[19] Mortazavi R, Jalili S. Enhancing aggregation phase of microaggreg-ation methods for interval disclosure risk minimization [J]. Data Mining and Knowledge Discovery, 2016, 30 (3): 605–639.